头豹 LeadLeo

行业简报

DeepSeek对AI产业的影响

深度分析DeepSeek爆火背后，AI产业将面临怎样的颠覆与冲击？

企业标签：深度求索、幻方量化、OpenAI

AI变革行业创新发展

China Large Model Industry

中国ビッグモデル産業

撰写人：陈庆民

团队介绍

头豹是国内领先的行企研究原创内容平台和创新的数字化研究服务提供商。头豹在中国已布局3大研究院，拥有近百名资深分析师，头豹科创网(www.leadleo.com)拥有20万+注册用户，6,000+行业赛道覆盖及相关研究报告产出。

头豹打造了一系列产品及解决方案，包括数据库服务、行企研报服务、微估值及微尽调自动化产品、财务顾问服务、PR及IR服务，研究课程，以及分析师培训等。诚挚欢迎各界精英与头豹交流合作，请即通过邮件或来电咨询。

报告作者

袁栩聪
首席分析师
oliver.yuan@Leadleo.com

陈庆民
行业分析师
qingmin.chen@leadleo.com

头豹研究院

咨询/合作

网址：www.leadleo.com

电话：15999806788（袁先生）

电话：13080197867（李先生）

深圳市华润置地大厦E座4105室

名词解释

- **蒸馏：**在AI开发中，通过使用更大能力强的模型的输出，在较小的模型上获得更好的性能的技术。这使得开发者能够以更低的成本在特定任务上获得类似的结果。
- **黑箱隔离（Black Box Isolation）：**一种技术手段，指的是即使不直接访问更大的模型，也能通过大量的查询和结果分析来改进较小模型的性能。
- **API（Application Programming Interface）：**应用程序接口，是一组定义、程序和协议的集合，允许不同的软件应用程序之间进行交流。
- **推理成本（Inference Cost）：**指运行AI模型以生成预测或决策时所需的计算资源成本。与训练成本相比，推理成本通常涉及持续性的运算消耗。
- **GPU（Graphics Processing Unit）：**图形处理单元，是一种专门用于加速图像渲染以及并行计算任务的处理器。
- **稀疏度（Sparsity）：**指神经网络中参数为零的比例。高稀疏度意味着大多数参数为零，可以减少计算量和存储需求。
- **PTX（Parallel Thread Execution）：**NVIDIA CUDA架构的底层虚拟机指令集，旨在提供跨不同代际NVIDIA GPU的兼容性编程层。
- **FP8（Floating Point 8-bit）：**一种8位浮点数格式，用于降低内存占用和计算需求，同时保持一定的数值精度。
- **CUDA（Compute Unified Device Architecture）：**由NVIDIA开发的一个并行计算平台和编程模型，使开发者可以利用NVIDIA GPU执行通用计算任务。

Chapter 1
DeepSeek
对AI大模型产业影响

- DeepSeek低训练成本影响
- DeepSeek引发的全球开源闭源模型讨论
- DeepSeek对英伟达算力市场的冲击与重构
- DeepSeek引发的模型蒸馏与知识产权的争议

DeepSeek的低成本训练模型对AI行业有何影响？

- DeepSeek的成本争议不仅关乎成本数字，而是预示着行业范式的迁移，即推理效率取代训练规模成为AI商业化的核心瓶颈。大模型赛道的最终赢家将是那些在“算法效率”与“成本透明度”上双赢的企业

DeepSeek V3训练成本争议背后的技术范式与行业博弈分析

事件背景

DeepSeek引发的争议主要集中在其低成本的训练模型上，尤其是V3模型的训练费用。DeepSeek的V3模型训练仅花费了557.6万美元（通过租赁278.8万个H800 GPU小时来计算的成本，平均每小时租金为2美元折算），大概是GPT-4的1/20。

1 *DeepSeek强调的“一次性训练成本”与硅谷看重的“全周期开发成本”碰撞出成本统计口径"罗生门"现象*

一次性训练成本统计口径支持方

一次性训练成本统计口径的支持方通常认为，DeepSeek之所以能实现低训练成本，得益于云计算资源的灵活租赁。**与传统自建集群模式相比，DeepSeek通过租赁公有云中的GPU，减少了对固定硬件的依赖，进而减轻了初期硬件投资的压力。这种做法更契合精益创业思维，即以较小成本进行快速实验和迭代，在一定程度上降低风险。**精益创业思维强调在资源有限的情况下快速创新，利用弹性计算优化硬件资源使用，而非一开始就大规模投入资本。支持方认为，这种模式挑战了传统硬件采购模式，能在短期内实现技术突破和获得竞争优势。

全周期开发成本支持方

全周期开发成本的支持者强调，在传统大型模型开发中，必须重视整个生命周期中的成本投入。其核心观点是，**研发过程中诸如硬件采购和试错成本等投资不仅不可忽视，而且应被视为“沉没成本资产化”的一部分。**例如，DeepSeek仅在硬件采购上可能就花费了超过5亿美元。此外，DeepSeek开发新的架构（如MLA架构和稀疏模型）通常需要数月的时间进行实验、调优和验证，期间失败尝试带来的成本同样非常高昂。因此，DeepSeek所公布的557.6万美元训练成本严重低估了实际总投入，未能全面考虑包括硬件采购、人员薪酬及试错成本在内的全周期开发成本。

2 *成本统计口径争议的本质为东西方技术路线的范式冲突*

当前AI技术竞争中，成本争议并非仅仅局限于数字的高低，背后其实蕴含着东西方技术路线的深刻分歧。这种分歧不仅反映在AI模型的研发成本上，更体现在两种截然不同的商业模式和技术发展思维中。

美国的“重资产投入”模式

美国在人工智能技术研发领域呈现出显著的资本密集型发展路径，其战略选择体现了对技术主导权的系统性布局。以NVIDIA为代表的半导体巨头为例，其研发的A100、H100等高性能计算芯片已形成全球AI基础设施的算力支柱，这种技术优势的构建源于多重战略要素的叠加，例如依托雄厚的研发投入（NVIDIA公司在2024年的研发支出为86.75亿美元，比2023年增长了18.2%）、尖端硬件基础设施的持续迭代，以及资本密集型科技企业的生态协同等。通过战略性的资本运作和技术资本化路径，美国科技企业能够快速形成技术代际差，建立产业链关键节点的市场支配地位，继而通过技术授权、云服务平台和算力租赁等多元商业模式实现技术红利的指数级变现。

来源：头豹研究院

（接上页——低成本训练模型对AI行业影响）

■ **中国的“轻量化架构”模式**

中国AI公司正在积极探索轻量化、高效迭代的技术路径，与西方重资产模式构成战略分野。**中国AI公司倾向于设计轻量化架构，减少硬件依赖，通过快速迭代技术来提高开发效率。**此方式不依赖高成本硬件采购，而是侧重于技术创新与软件优化以降低成本。**该发展模式的核心价值在于，以技术密度替代资本密度，在半导体技术代差存在的情况下，构建从技术突破到商业变现的“高速通道”。**

■ **硬件采购成本的隐藏议题**

目前，全球AI训练芯片市场由英伟达主导，其A100和H100系列凭借强大的CUDA生态系统，占据了市场份额的90%以上。英伟达不仅在性能上处于领先地位，更通过构建一个完善的软件生态系统，进一步巩固了其在行业中的优势地位。随着美国政府对中国的AI芯片出口实施禁运，特别是针对A100和H100等核心芯片，中国AI公司在硬件采购方面面临着前所未有的挑战。为了应对这一困境，中国AI企业不得不寻求替代方案，迅速转向国产算力，并加速构建“去英伟达化”的技术生态。例如，华为的昇腾910C芯片，其性能可达到H100的60%，成为中国AI公司在芯片供应链受到限制时的重要选择。

3 *DeepSeek的技术路径揭示了大模型赛道一个关键趋势：从“训练军备竞赛”转向“推理效率革命”*

传统上，AI行业的资源投入主要集中在模型训练阶段，但随着推理效率的提升和成本控制的突破，未来的行业发展将更加聚焦于推理阶段的优化。这一趋势不仅推动了技术创新，也为行业带来了新的商业机会和竞争态势。

■ **技术逻辑：稀疏架构与推理效率的“降维打击”**

参数激活率决定长期成本

在传统的大型AI模型中，如GPT-4，推理过程需要激活模型中的所有参数。随着模型规模的不断扩大，推理时所需的计算资源和算力呈线性增长，这导致了训练和推理过程中的成本不断上升。尤其是在大模型的商业化应用中，推理阶段的成本成为企业持续运营的重要负担。

与传统做法不同，DeepSeek通过采用稀疏架构，使得在推理时仅激活模型中的35-37%的参数。这种技术创新显著提高了推理效率，缩短了单次推理的时间，同时提高了单位GPU的吞吐量。由于稀疏架构的引入，DeepSeek实现了推理时的“降维打击”，即在同等硬件资源下，能够处理更多的计算任务，从而降低了长期推理成本。

技术外溢效应：可推动行业标准

如果MLA（潜在注意力机制）等稀疏架构成为行业标准，那么AI模型优化的核心指标将从传统的“参数量”转向“激活效率”。这种变化将推动更加轻量化和场景专用的模型崛起。这些模型不仅能够高效执行特定任务，还能在边缘计算等对计算资源有较高要求的场景中广泛应用，进一步推动AI技术的普及和商业化应用。

■ **商业逻辑：从“烧钱训练”到“订阅式服务”的盈利重构**

成本结构的颠覆

传统的AI行业商业模式通常依赖于高昂的训练成本，这些成本尽管是一次性投入，但也往往在企业的初期投入中占据了很大一部分。然而，**推理成本**因其长期性和持续性，成为了AI企业持续经营的负担。据估算，推理成本可能占到企业AI支出总额的70%以上。随着DeepSeek在推理效率上的突破，其模型不仅能够降低单次推理的时间和算力需求，而且能够在**长期服务**中实现显著的成本节约。

这种推理效率的提升，直接降低了企业的**服务边际成本**，使得企业能够更容易实现规模化盈利，特别是在**按调用次数收费**的API模式下，企业可以通过更低的成本提供高效的AI服务，从而降低了盈利门槛。

来源：头豹研究院

（接上页——低成本训练模型对AI行业影响）

应用层的爆发机遇

随着推理成本的下降，AI技术在多个高频场景中的应用将得到加速，尤其是在边缘计算和实时交互（如智能客服、游戏推荐等）等领域。低成本、高效率的推理技术使得AI能够在**场景适配性**更强的情况下，快速渗透到各种业务场景中，催生出新的商业模式。这些新兴应用场景不仅扩展了AI技术的商业化空间，还推动了新的盈利模式的产生，如按使用量收费、基于性能的付费等。

■ **行业竞争格局：头部玩家壁垒松动，中小厂商迎逆袭机会**

训练成本门槛的降低

传统的大模型开发依赖于庞大的硬件资源和数据支持，这使得AI技术的领先地位通常由少数几家技术巨头主导，如OpenAI、Google等。然而，随着稀疏架构的开源和技术的不断扩散，中小厂商有机会基于有限的算力进行高性能模型训练。这一趋势将打破传统的大型企业在“**数据+算力**”上的垄断，为新兴企业提供逆袭的机会。中小厂商可以借助更轻量化的模型架构，迅速实现技术突破，降低AI训练的高门槛。

差异化竞争的焦点

未来的AI行业竞争将不再仅仅集中在参数量的竞争上，更多的焦点将转向“激活效率”和“场景适配度”。随着推理技术的进步，AI模型将不再单纯地依赖庞大的参数规模，而是依赖于如何在具体场景中优化模型的性能，提升资源的利用率。数据质量、算法轻量化能力和场景的适配性将成为新的技术竞争护城河，行业竞争将更加注重“**技术差异化**”和“**服务定制化**”。

4 *大模型成本争议背后的深层启示为AI行业需要“成本透明度标准”*

当前AI行业关于DeepSeek与其他企业成本数据披露的争议焦点，根本上揭示了行业内成本透明度的缺失。这不仅影响到行业的公信力和竞争公平性，更可能延缓技术创新和商业模式的健康发展。在AI领域，**缺乏统一的成本核算标准**，使得不同企业在成本数据的披露上存在巨大差异。尤其是一些企业选择**只公开训练成本**而隐藏了更为庞大的**硬件采购**和**试错投入**，导致外界难以全面评估其研发投入和技术能力。因此，推动AI行业建立**明确的成本透明度标准**，已成为行业未来发展的关键议题。

■ **建立全生命周期成本披露框架**

当前，AI行业普遍存在“成本披露碎片化”现象，即企业可能仅披露部分成本数据（如训练成本），却忽略了硬件采购、试错过程等隐性开支，进而使得外界对其技术和商业模式的判断受到偏差。为此，建立**全生命周期成本披露框架**显得尤为重要。该框架应覆盖AI模型的每一个开发和运营阶段，从而为外界提供更为全面和透明的成本结构。

❑ **AI大模型各阶段成本的组成**

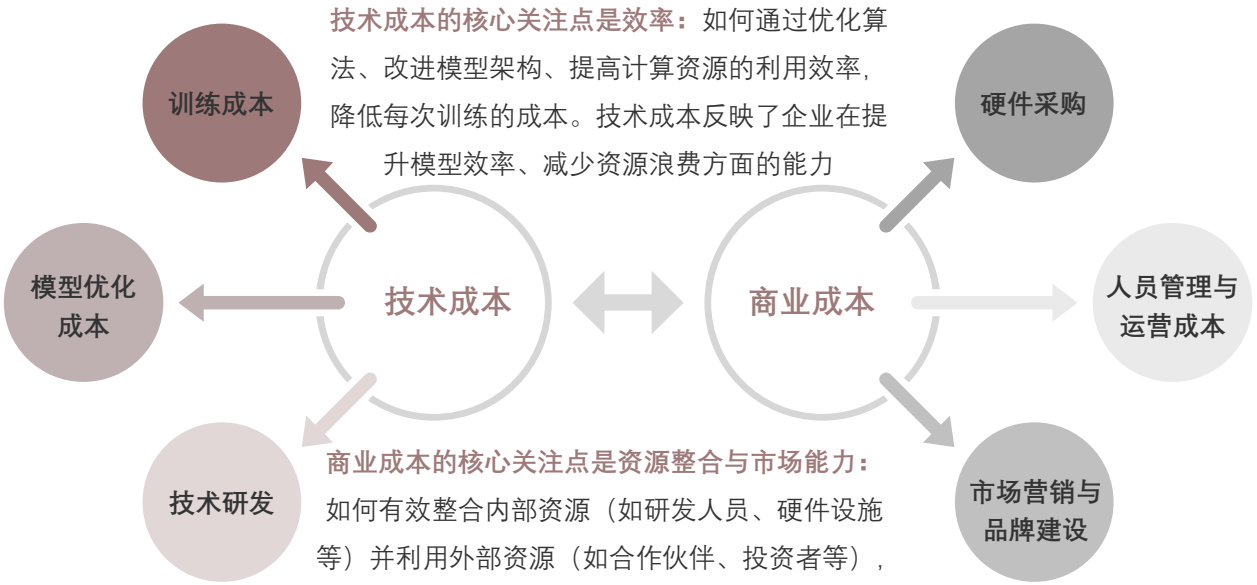
硬件购置/租赁成本	算法开发与试错迭代成本	部署与运维成本
企业能依自身情况选购买硬件或租赁GPU资源，二者成本结构差异显著。构建透明硬件成本披露标准，可有效防止企业公开数据时只呈租赁成本或忽略采购费用	AI模型开发除初期的训练数据和硬件，还涉及大量的实验和迭代，如开发新算法和调整模型结构，这些都需大量的GPU计算时间和人工。这部分成本是评估AI模型“技术门槛”的关键部分，应披露在成本中	AI技术不止于模型训练，还需场景部署与运维，涵盖服务器、存储、网络等成本。企业初期开发常忽视长期运维支出，致长期成本低估，应明确这些持续成本

来源：头豹研究院

（接上页——低成本训练模型对AI行业影响）

■ **区分“技术成本”与“商业成本”**

AI行业的成本结构十分复杂，因此，必须明确区分“技术成本”与“商业成本”，并根据不同维度进行详细分析，以全面评估企业的竞争力和市场潜力。



区分这两者有助于从不同角度评估企业的技术优势和商业潜力。在AI行业中，技术成本和商业成本之间的平衡决定了企业能否在竞争激烈的市场中脱颖而出。

■ **推动行业透明度：增强投资者信任与市场公平**

建立成本透明度标准不仅有助于企业展示其技术和运营能力，还将增强投资者的信任。投资者在评估AI企业时，往往依赖于成本数据来判断企业的技术成熟度和长期盈利能力。没有透明的成本核算，投资者可能会错失对新兴企业价值的准确评估，从而影响行业资本流动的效率。

此外，透明的成本披露也有助于推动市场的公平竞争。在AI行业的早期阶段，**“数据-算力”垄断**往往是技术巨头的竞争优势，而缺乏统一的成本标准使得新兴企业难以有效展示其成本控制能力。通过建立统一的成本披露框架和标准，更多中小企业将能够清晰地展示其技术实力，打破市场的垄断局面，推动行业的多元化发展。

来源：头豹研究院

DeepSeek引发全球对开源与闭源模型的大讨论

- 大模型市场竞争激烈，技术优势效益正在递减，厂商聚焦成本优化。开源生态以长尾创新等改变下游采购逻辑，使闭源溢价缩至高精尖场景。同时，闭源商调整模式，借定制等重构价值定位

DeepSeek作为开源模型的代表对全球开源及闭源模型带来的影响分析

事件背景

2025年初，DeepSeek的技术成果和市场反响席卷全球，尤其是在美国、印度等重要市场，用户下载量和日活跃度激增，其技术突破让曾经以为自己处于领先地位的硅谷科技公司，尤其是OpenAI等闭源模型提供商感受到了巨大的压力。DeepSeek不仅在性能上与业内领先的闭源模型并肩，而且其开源的商业模式和价格优势使得开发者和公司对于闭源模型的依赖性产生了质疑。

1 *价格战正重塑大模型市场竞争，闭源技术优势的边际效益递减，加速市场开源与闭源格局重构*

在DeepSeek的崛起过程中，闭源模型，特别像OpenAI这样的公司，面临的最大挑战是价格战。闭源模型通常凭借其独特的技术积累和品牌溢价，长期占据市场的主导地位。然而，DeepSeek的开源模式则为开发者和企业提供了一个性能相当但价格更为低廉的替代方案，这直接冲击了闭源模型的商业模式。

价格战正重塑大模型市场竞争

当下大模型市场迈入“性能 - 成本”双维竞争期。DeepSeek - V2借架构创新削减大模型尤其是推理成本，API定价每百万tokens输入1元、输出2元（32K上下文），仅约OpenAI GPT - 4 Turbo价格的1%。此结构性价差瓦解了传统技术溢价模式，并且，开源模型在代码生成等部分场景性能超头部闭源产品。市场开始依循半导体“安迪 - 比尔定律”迭代，即，性能遇边际效应时，成本优化成关键竞争维度。这种转变迫使闭源厂商必须在技术突破和成本控制间建立新的平衡公式。

2 *开源生态在边缘端的“长尾创新效应”将形成技术民主化浪潮，动摇闭源模型的价值基础*

开源模型挤压闭源模型商业空间

开源社区通过Llama.cpp等优化框架，在推理效率上取得了指数级的提升。例如，70B的大型模型可以在树莓派设备上顺利运行；AirLLM更是展示了在仅有4GB显存的GPU上成功运行70B级别的Qwen模型，甚至能在8GB显存的设备上运行405B的Llama3.1。这些案例有力地证明了边缘端部署的可行性。随着技术的不断民主化，企业的采购决策逻辑也在发生根本变化：当开源方案能够满足80%的基础需求时，闭源产品的溢价空间便只能局限于那些极其高精尖的应用场景。

3 *闭源厂商被迫启动商业模式“三重转型”，AI大模型行业进入价值重构深水区*

定价策略转型：降低价格以应对开源模型的竞争优势

随着开源模型的出现，价格优势成为了开发者和企业的重要考量因素。闭源厂商，如OpenAI，在意识到价格过高将直接导致市场份额流失后，开始进行降价调整。以OpenAI为例，其发布的O3 mini模型在定价上比之前的O1 mini大幅降低超过60%。这一举措反映了闭源厂商对开源模型价格优势的直接回应。然而，降低价格的背后也带来了巨大的盈利压力，因为闭源模型的研发、训练以及基础设施成本通常极为昂贵。以Meta和Google为代表的公司，尽管在推理阶段能够实现高利润，但在前期的大规模训练投入和基础设施支出上，仍然面临严峻挑战。

来源：头豹研究院

（接上页——开源与闭源模型大讨论）

■ **开源或开放部分功能：提高技术竞争力同时保护核心技术**

面对开源模型带来的技术挑战，许多闭源公司开始探讨将部分功能或数据集开放的可能性。通过开源或部分开源，闭源厂商能够加速技术创新，提升产品竞争力，但同时也需要在开放与保护企业核心技术之间找到平衡点。开放部分功能既可以增强市场吸引力，也能促进开发者社区的参与，但核心技术的保护仍然是一个关键问题。例如，微软在其AI产品中采用了混合开源策略，包括Azure AI平台和基于GPT架构的语言模型。为了应对市场挑战，微软开源了Phi-3 Mini这样的小型语言模型，旨在展示其技术实力并吸引更多用户。这不仅增强了微软产品的吸引力，还促进了开发者社区的参与，使更多开发者接触到微软的技术栈，并可能转化为未来的付费用户或贡献者。同时，微软依然对其核心技术如更高级别的GPT系列模型保持严格的闭源控制，仅通过订阅服务提供给企业客户使用。这种方法确保了微软及其合作伙伴能够在竞争中保持领先地位，同时从高端产品中获得经济收益。

■ **商业模式多元化：通过增值服务和定制化解决方案实现收入多元化**

在面临开源模型竞争压力的背景下，闭源厂商不再仅仅依赖基础模型产品作为单一的收入来源，而是通过差异化的增值服务、API接口、定制化解决方案等方式，构建多元化的收入渠道。以OpenAI为例，其推出的企业专属微调服务将技术溢价从基础模型的销售转向了定制化能力的提供。这一转型实质上是在重构其价值定位：从传统的“卖模型”转向“卖服务深度”，从单纯的通用计算能力转向针对具体场景的智能解决方案。

4 *硬件-算法协同优化开启”后摩尔时代”成本革命，倒逼全产业链升级*

■ **硬件性能的提升将通过硬件与算法的深度协同来实现突破**

在“后摩尔时代”，硬件性能的提升不再依赖传统的摩尔定律，而是通过硬件与算法的深度协同来实现突破。例如，NVIDIA的H100到B200能效比在四年内提升了15倍，而在此期间，模型参数的增长速度却高达1000倍，这一“剪刀差”迫使行业寻找全面的创新解决方案。随着硬件性能的提升，单纯依靠硬件算力或算法本身的传统模式已经无法满足日益复杂的需求。

以美国Groq公司为例，该公司在其LPU（Layer Processing Unit）芯片上成功实现了与DeepSeek结合的高效推理，效率比最新的NVIDIA H100快了一个数量级，达到了每秒24,000个token。这表明，单纯依赖硬件算力的堆砌已不再是推动人工智能发展的唯一途径。随着芯片制造工艺逐渐接近瓶颈，未来的人工智能进步将主要依赖算法优化和芯片架构的协同创新。

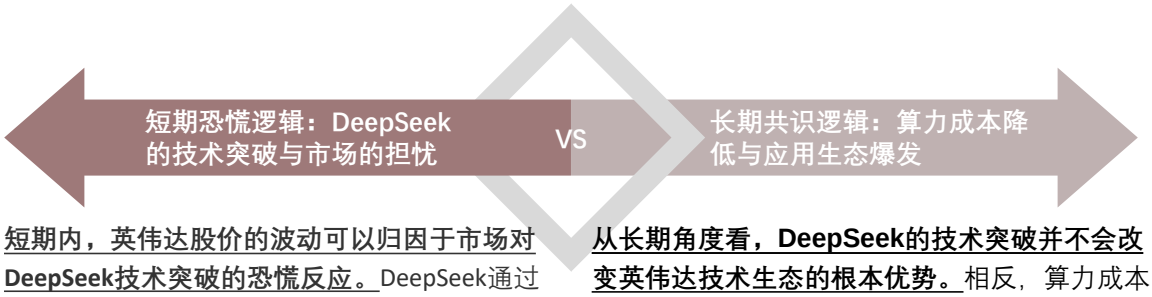
在这一全栈式创新浪潮的推动下，单一技术环节的竞争优势已经不再持久，其技术生命周期（半衰期）已缩短至12至18个月。因此，企业若要在这场成本革命中占据领先地位，必须构建一个从芯片到应用的完整技术生态。在这种生态体系中，硬件和算法的紧密协同将释放出更大的创新潜力，推动整个产业链的升级和重构。

来源：头豹研究院

DeepSeek技术突破对英伟达算力市场的冲击与重构

- 通过优化PTX层和支持FP8计算，短期内对英伟达构成竞争压力，但并未真正突破CUDA生态的核心优势。随着算力成本降低和开源生态发展，AI大模型算力市场格局将逐步分化

DeepSeek对英伟达算力市场带来的短期情绪与高端算力长期需求底层逻辑的博弈

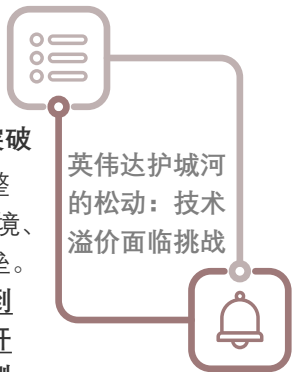


短期内，英伟达股价的波动可以归因于市场对DeepSeek技术突破的恐慌反应。DeepSeek通过优化PTX层和支持FP8计算，理论上减少了AI模型训练对英伟达高端GPU的依赖。PTX（Parallel Thread Execution）作为NVIDIA CUDA架构的底层指令集，长期以来构成了英伟达技术壁垒的核心部分。DeepSeek的突破主要集中在PTX层的局部优化，提升了计算效率，但其技术的另一个潜在影响是，DeepSeek是否也能具备在其他GPU平台（非最精尖）上进行优化的能力，这意味着其他厂商的GPU产品可能会迎头赶上，形成与英伟达产品的竞争局面。

从长期角度看，DeepSeek的技术突破并不会改变英伟达技术生态的根本优势。相反，算力成本的降低将推动AI应用生态的爆发，进而为英伟达带来更多的市场需求。AI算力的成本降低将使得更多中小型企业能够参与到AI应用开发中，推动包括开源模型复现和推理场景在内的应用场景的扩展。例如，推理计算的下沉和优化将使得AI技术可以更广泛地应用于边缘计算、物联网等领域，这将大大增加对算力的需求。随着算力的需求不断增加，英伟达等领先企业将面临技术迭代与创新的压力，不断优化GPU架构和算力产品，以适应日益增长的市场需求。

关键矛盾点：市场是否存在误判DeepSeek技术的破坏性：本次波动的关键矛盾点在于，市场对DeepSeek技术突破的“破坏力”存在误判。DeepSeek的突破并非如一些市场评论所言，能完全绕过英伟达的技术护城河。而是在CUDA生态的底层架构上进行了局部优化，特别是在PTX指令集的改进上。这种改进类似于在现有操作系统（如Windows）上开发更高效的编译器，而非从根本上改变操作系统本身。

英伟达护城河的“松动”与“加固”



生态分层突破

英伟达不仅在硬件上占据优势，其完整的硬件与软件生态（包括CUDA编程环境、开发工具和支持库）构成了其技术壁垒。然而，**随着DeepSeek的模型逐步集成到其他厂商的生态中，像AMD等公司也开始尝试在特定领域，尤其是推理计算侧，建立自己的替代生态。**这种生态的分层突破，意味着英伟达的“全栈垄断”模式正面临冲击。

FP8计算与内存优化

DeepSeek的出现标志着低精度计算（如FP8）的突破，理论上能有效降低训练成本，提升计算效率。FP8的支持使得在某些特定场景下，AI模型训练可以不依赖于英伟达高端GPU，转而采用性能较为中低端的GPU或专用芯片，如谷歌的TPU（张量处理单元）。这为其他厂商提供了替代英伟达硬件的可能性，逐步削弱了英伟达在高端算力市场的溢价能力。随着FP8计算的推广，越来越多的AI应用场景能够适应中低端硬件，这意味着从事AI研发的公司将不再局限于英伟达高端GPU的选择。

来源：头豹研究院

（接上页——对英伟达算力市场的冲击与重构）

护城河的加固：生态粘性与技术纵深

CUDA生态的不可替代性

尽管DeepSeek技术在局部优化上取得了进展，但CUDA生态依然是英伟达不可替代的优势。CUDA不仅仅是一个API接口，它还涵盖了从编译器、库函数到开发者社区的全生命周期支持。这个完整的生态体系是英伟达的技术壁垒之一，也是许多开发者和企业选择英伟达硬件的重要原因。即便在硬件层面，竞争者能够推出优化的计算路径，绕过高层的API，但底层的PTX指令集依然依赖于英伟达的GPU架构，其他厂商无法轻易复制这一点。

芯片互联技术的壁垒

英伟达在芯片互联技术上的优势，特别是其NVLink和InfiniBand技术，依然是竞争对手难以超越的技术壁垒。NVLink通过提供更高的带宽和更低的延迟，使得多个GPU之间可以高效协同工作，在大规模AI训练中展现出无可匹敌的优势。而InfiniBand则为大规模数据中心提供了高速数据传输的解决方案。这些技术在多GPU集群的训练中提供了极大的效率提升，使得英伟达在超算领域依旧占据领先地位

AI算力市场竞争格局的路径重构分析

短期格局：多元竞争与成本分层

推理市场分化

在AI推理领域，随着DeepSeek等新兴技术的崛起，算力需求开始分化。谷歌的TPU、AMD的Instinct GPU、AWS的Trainium等专用芯片，尤其在标准化推理场景中，逐渐成为英伟达的强有力竞争者。这些专用芯片凭借其高效的计算性能和优化的推理能力，能够在推理任务中获得更多市场份额，进而对英伟达的显卡产品构成挑战。

训练市场仍为主战场

尽管推理市场日益分化，但大模型训练仍将是算力市场的主战场。大规模模型训练对算力的要求极高，尤其依赖于多GPU互联和CUDA工具链。英伟达的H100和A100 GPU依然是训练市场中不可或缺的“刚需”产品。当前，许多领先的AI研发和企业级应用仍然依赖英伟达的高端GPU来满足其庞大的计算需求。

长期变量：开源生态与标准化博弈

开源社区的“双刃剑”效应

随着DeepSeek等技术的推动，开源AI模型和平台成为行业趋势，降低了开发者和研究人员的接入门槛，并逐步解决了硬件平台兼容性问题。如果DeepSeek推动更多模型开源，将为跨硬件平台的标准化接口奠定基础，类似于PyTorch对深度学习框架的影响。

英伟达的应对策略

面对开源生态和标准化接口的发展，英伟达将通过开放部分生态（如推出兼容AMD GPU的CUDA Lite版本）来增强竞争力，减少对单一硬件的依赖，并扩大影响力。此外，**收购关键开源项目或与开源社区合作也可能成为战略选择。**

来源：头豹研究院

（接上页——对英伟达算力市场的冲击与重构）

英伟达投资视角下的关键信号

英伟达的“压力测试指标”

毛利率变化：

毛利率是衡量公司盈利能力的重要指标，尤其在硬件产业中尤为敏感。英伟达目前的毛利率约为74.6%，这一数字体现了公司在高端GPU市场的定价优势和技术溢价。如果在接下来的2-3个季度，英伟达的毛利率出现显著下降，尤其是跌破65%，这将是一个警示信号，表明市场竞争加剧，价格战已经实质性开始。具体来说，价格战可能由多个因素驱动，如：新兴竞争者推出性价比更高的产品，或者云服务商和大型数据中心开始通过自研芯片降低对英伟达的依赖。在这种情况下，英伟达将不得不通过降价来维持市场份额，进而影响整体盈利能力。

客户结构迁移：

英伟达的客户结构中，超大规模云厂商（如AWS、谷歌、微软等）占据重要地位。这些云厂商长期以来依赖英伟达的GPU产品来支持其AI和大数据计算需求。然而，随着这些云厂商逐渐加大自研芯片的投入，若其自研芯片的比例均提高至50%以上，这将对英伟达构成长期威胁。自研芯片不仅可以降低对外部硬件的依赖，还能通过更好地优化与自身基础设施的兼容性，提升整体计算效率和成本效益。

颠覆性技术突破的观测点

PTX层通用化工具链：

PTX是英伟达CUDA架构的核心组成部分，长期以来作为英伟达GPU的技术壁垒。如果出现类似LLVM（低级虚拟机）的跨厂商中间表示层（IR），这一技术突破可能会对CUDA生态造成重大冲击。LLVM作为一个跨平台的编译器框架，允许不同硬件平台之间共享中间代码表示，极大地降低了对特定硬件（如英伟达GPU）的依赖。如果这种跨厂商的通用化工具链成为行业标准，开发者将能够更轻松地在不同厂商的硬件上部署AI计算任务，进一步削弱CUDA和英伟达硬件的绑定效应。对此，投资者应密切关注相关技术是否快速发展并获得广泛支持。若这一趋势加速，可能意味着英伟达的技术优势会在未来数年内逐渐淡化，从而影响其市场地位和盈利模式。

光计算/存算一体芯片：

光计算与存算一体化芯片（即计算与存储单元在同一芯片上集成）是新型计算架构中的一项前沿技术。该技术通过利用光学信号进行信息传输，能够突破传统电子计算的瓶颈，尤其是在能效方面展现出巨大的潜力。如果这种技术在特定场景下，如稀疏模型训练，能够实现10倍以上的能效比突破，将会彻底重构算力竞争的维度。这种突破性技术不仅能够大幅提升计算效率，还可能会大大降低算力成本，尤其在数据密集型的AI应用中，具有极大的市场潜力。投资者需要关注这一技术的进展，特别是技术实现的难度、商业化落地的速度以及对传统硬件厂商（如英伟达、AMD等）带来的潜在影响。如果光计算技术或存算一体芯片的研发进程加速，可能会带来一场新一轮的算力革命，迫使现有的硬件厂商进行战略调整，甚至颠覆现有的市场格局。

来源：头豹研究院

DeepSeek引发的模型蒸馏与知识产权争议

- OpenAI指控DeepSeek通过未经授权的API查询，利用其模型输出数据进行模型蒸馏，构成知识产权侵权。这类指控的关键在于证明数据来源及使用是否违反服务协议条款，尤其是API使用的限制等

模型蒸馏技术是否构成知识产权侵权？法律界如何界定此类争议？

事件背景

OpenAI指控DeepSeek通过“蒸馏”技术使用其数据训练模型，涉嫌侵犯知识产权

模型数据蒸馏技术解释

模型蒸馏本质上是将一个复杂的、高性能的生成的数据或输出用于训练一个较小、效率更高的模型。此过程通常会保留大模型的部分知识和特性，但并不需要直接访问大模型的内部结构或代码。这种技术在学术界和行业中都被视为一种有效的手段，能够在保证较低成本的同时，获取类似大模型的性能。

1 *AI大模型界使用的蒸馏技术是否在现行法律下构成知识产权侵权*

在这类争议中，关键问题在于是否存在未经授权的使用和数据窃取。例如，OpenAI指控DeepSeek在训练模型时，可能利用了OpenAI的模型生成的输出数据，通过大量查询OpenAI的API进行“黑盒操作”，从而“蒸馏”出其模型的表现。这种操作可能导致DeepSeek的模型性能与OpenAI的模型相似，因而构成了侵权的证据

技术层面：蒸馏是利用已有大模型的知识输出来训练小模型。这种方式本身并不违法，但若没有适当授权或使用公开数据集外的私有数据，那么就是违法的。例如，若DeepSeek通过使用OpenAI的API且未遵循OpenAI的服务协议进行模型训练，那便可能涉及到知识产权侵权的问题。

- **API服务的使用条款**：OpenAI的API服务通常会有严格的使用条款和授权协议，明确规定了如何使用其返回的数据。例如，OpenAI可能规定，用户只能将API查询结果用于个人用途、非商业性应用，或在某些情况下，禁止将这些数据用于开发与OpenAI产品竞争的技术。
- **未经授权使用API数据**：如果DeepSeek在没有遵守OpenAI的授权协议的情况下，通过大量查询OpenAI的API并收集其返回的数据进行模型训练，那么这就有可能违反OpenAI的使用条款，构成**未经授权的使用**。
- **黑盒操作：**用户通过调用API并获取其输出，但不直接接触或理解API背后的源代码、训练数据或模型参数。这种操作有时被用来“反向工程”出API模型的性能特点。即，虽然用户不直接访问API模型的核心数据或算法，但通过大量的输入-输出交互（如频繁查询API并获取响应），可以间接推测出模型的内部行为和结构，从而**复制模型的表现**。

法律界的界定标准

数据使用的合法性	服务协议条款	竞争对手的界定
如果使用了受版权保护的训练数据或模型输出而未获得授权，可能构成数据盗用	OpenAI的服务协议中可能明确规定了其模型输出的使用限制，禁止将其数据用于开发竞品	如何定义“竞争对手”通常是一个模糊点，具体是否构成侵权取决于法律如何界定直接竞争

来源：头豹研究院

（接上页——模型蒸馏与知识产权争议）

OpenAI指控DeepSeek侵权的法律分析

■ 侵权指控的核心与法律基础

OpenAI指控DeepSeek在训练其模型时，通过频繁查询OpenAI的API并利用其返回的数据，进行“黑盒操作”以复制OpenAI模型的输出。根据OpenAI的服务协议，明确禁止将其生成的数据用于开发与其竞争的技术产品，尤其是通过蒸馏技术提取大模型的知识来训练小模型，构成违反服务协议的行为。具体而言，服务协议中的限制条款保护了OpenAI的知识产权，尤其是模型输出数据的使用，这类数据通常代表着OpenAI独特的技术积累与创新。因此，若DeepSeek确实通过未授权的API查询，利用这些输出数据进行模型训练，便可能构成违约行为。

然而，值得注意的是，**侵权指控的成立必须依赖于明确的证据，而这正是本案中的关键法律难点。OpenAI需要证明DeepSeek确实通过其API获取了足够的专有数据，并在此基础上进行模型蒸馏。简单的“相似性”或“模仿”并不足以成立侵权，需要具体的数据证据来支撑这一指控。**

■ 举证难点：如何证明数据的直接来源

法律争议的核心问题在于举证责任。在此类案件中，原告方需要证明被告确实获取并利用了专有数据，而非通过其他公开渠道或独立研发路径获得了类似的结果。OpenAI若要证明DeepSeek侵犯其知识产权，首先需要提供**直接证据**，例如API调用记录或日志，这些可以清晰地表明DeepSeek是通过OpenAI的API生成的具体数据来训练其模型。这类证据若能确认数据的源头及其使用方式，则有可能构成对侵权行为的有力支持。

然而，在缺乏足够证据的情况下，这类指控通常会使法院难以认定这种“间接复制”的行为。类似的案例，如纽约时报诉OpenAI案，法院在审理过程中明确要求原告提供具体的复制比例或独特输出模式，而不是仅凭“相似性”或“行为模式”作出判断。

■ 行业规范与法律滞后

在学术界，模型蒸馏通常被视为一种合法且常见的做法，目的是提升模型性能并降低计算成本。但当这种做法在商业环境中被用作获取竞争优势时，涉及的法律问题便变得复杂。**目前，AI行业在大模型的开发与数据使用方面缺乏明确的法律框架和行业规范。大模型的“训练数据”与“输出数据”之间的边界模糊，许多公司在没有明确法律指导的情况下进行操作，导致了法律争议。**

来源：头豹研究院

成为头豹会员—享专属权益

- 成为头豹会员，尊享头豹海量数据库内容及定制化研究咨询服务
- 头豹已累积上万本行业报告、词条报告，拥有20万+注册用户，沉淀100万+原创数据元素
- 头豹优势：行业覆盖全、数据量庞大、研究内容应用场景广泛，并有专业分析师团队为您提供定制化服务，助力企业展业

报告次卡

任意10本报告
阅读权益（一年有效）

¥598 /年

企业标准版

适用于研究频次高的用户或企业
无限量阅读全站报告
升级报告下载量
专享企业服务
定制词条报告

¥50,000 /年

企业专业版/旗舰版

满足定制研究需求的企业用户
定制深度研究报告
随需下载报告
分析师一对一沟通
专享所有核心功能

¥150,000+ /年

购买与咨询

咨询邮箱：

nancy.wang@frostchina.com

客服电话：

400-072-5588

方法论

- 头豹研究院布局中国市场，深入研究19大行业，持续跟踪532个垂直行业的市场变化，已沉淀超过100万行业研究价值数据元素，完成超过1万个独立的研究咨询项目。
- 研究院依托中国活跃的经济环境，研究内容覆盖整个行业的发展周期，伴随着行业中企业的创立，发展，扩张，到企业走向上市及上市后的成熟期，研究院的各行业研究员探索和评估行业中多变的产业模式，企业的商业模式和运营模式，以专业的视野解读行业的沿革。
- 研究院融合传统与新型的研究方法，采用自主研发的算法，结合行业交叉的大数据，以多元化的调研方法，挖掘定量数据背后的逻辑，分析定性内容背后的观点，客观和真实地阐述行业的现状，前瞻性地预测行业未来的发展趋势，在研究院的每一份研究报告中，完整地呈现行业的过去，现在和未来。
- 研究院密切关注行业发展最新动向，报告内容及数据会随着行业发展、技术革新、竞争格局变化、政策法规颁布、市场调研深入，保持不断更新与优化。
- 研究院秉承匠心研究，砥砺前行的宗旨，从战略的角度分析行业，从执行的层面阅读行业，为每一个行业的报告阅读者提供值得品鉴的研究报告。

法律声明

业务合作

会员账号

可阅读全部原创报告和百万数据，提供PC及移动端，方便触达平台内容

定制报告/词条

行企研究多模态搜索引擎及数据库，募投可研、尽调、IRPR等研究咨询

定制白皮书

对产业及细分行业进行现状梳理和趋势洞察，输出全局观深度研究报告

招股书引用

研究覆盖国民经济19+核心产业，内容可授权引用至上市文件、年报

市场地位确认

对客户竞争优势进行评估和证明，助力企业价值提升及品牌影响力传播

云实习课程

依托完善行业研究体系，帮助学生掌握行业研究能力，丰富简历履历

业务热线

袁先生：15999806788

李先生：13080197867